특집논문 (Special Paper)
방송공학회논문지 제30권 제6호, 2025년 11월 (JBE Vol.30, No.6, November 2025)
https://doi.org/10.5909/JBE.2025.30.6.867
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

뉴럴 라이트필드 MLP 모델의 효율성 개선을 위한 Bit Precision 변화에 따른 성능 비교 연구

양 서 준[a], 정 현 민[a]‡

Performance Comparison according to Changes in Bit Precision to Improve the Efficiency of Neural Light Field MLP Model

Seojun Yang[a] and Hyunmin Jung[a]‡

요 약

딥러닝 모델의 복잡도, 데이터의 부담 절감은 딥러닝의 모바일 디바이스에서의 적용에 있어 필수적이다. 본 연구는 실감미디어에서 사용되는 뉴럴 라이트필드 (Neural Light Field)의 Multi-Layer Perceptron (MLP) 모델을 대상으로 비트 정밀도 조정에 따른 모델 크기 절감과 성능 저하를 분석한다. 본 논문은 MLP의 32bit 부동 소수점 형식의 가중치를 16bit, 8bit로 조정하며 성능을 비교한다. 16bit 일괄 조정 시 모델 크기를 절반으로 줄이면서 성능 저하가 거의 없음을 확인하였으며, 8bit 일괄 조정 시 기존 모델 대비 25%의 데이터 부담 절감을 보이나, 심각한 화질 저하가 관측되었다. 이에 은닉층별 민감도를 세밀하게 분석하고, 출력 layer에 가까운 은닉층들에 대하여 선택적으로 8bit로 조정하는 방법을 적용함으로써 화질 저화를 최소화하면서 데이터 부담 절감 효과를 높이는 방법을 제안한다.

Abstract

Reducing the complexity and data burden of deep learning models is essential for the deployment on resource-constrained mobile devices. In this study, we compare and analyze the trade-offs between model size reduction and performance degradation resulting from bit precision adjustments in the Multi-Layer Perceptron (MLP) model of the Neural Light Field, a widely used representation for immersive media. We adjust the single precision weights of the MLP to 16-bit and 8-bit. Experimental results show that models adjusted to 16-bit bit precision reduce their size by half with minimal image quality degradation. In contrast, models adjusted to 8-bit achieve an additional reduction in data burden of up to 25%, but suffer from significant image quality degradation. To address this, we analyze the sensitivity of each hidden layer and propose a method that applies 8-bit precision only to hidden layers close to the output. This approach effectively reduces data burden while minimizing image quality degradation.

Keyword : Implicit Neural Representation, Neural Light Field, Multi-Layer Perceptron

Ⅰ. 서 론

최근 확장현실 (eXtended Reality, XR) 기술이 게임, 의료, 교육 등 다양한 분야에서 혁신적인 사용자 경험을 제공하며 빠르게 성장하고 있다. 특히, Meta Quest[1], Apple Vision Pro[2] 등 다양한 XR 장비들이 상용화되고, 대중화됨에 따라, 몰입형 실감 콘텐츠에 대한 사람들의 관심과 기대가 더욱 높아지고 있다. 이러한 XR 생태계에서 자유 시점 뷰 합성 기술 (Free-viewpoint View Synthesis)은 핵심적인 역할을 담당한다. 사용자가 가상 공간 내에서 자유롭게 이동하며 원하는 시점에서의 고품질 뷰를 실시간으로 경험할 수 있게 하는 이 기술은, 전통적인 360도 파노라마 영상이 가지는 고정 시점의 제약을 넘어서 더 높은 몰입감을 제공한다[3].

최근 자유 시점 뷰 합성 기술은 Gaussian Splatting[4], Neural Radiance Field (NeRF)[5]의 등장으로 새로운 국면을 맞이하고 있다. Gaussian Splatting은 3D 장면을 모방하기 위해 3차원 공간 상에 다수의 3D Gaussian을 배치, 이의 위치, 크기, 회전 각도, 색, 투명도를 최적화하는 방법으로, 높은 화질과 빠른 속도를 자랑한다. NeRF는 연속적인 3D 좌표와 시점 방향을 입력으로 받아 Multi-Layer Perceptron (MLP)을 통해 해당 위치의 밀도 (density)와 색상 (color)을 예측하고, 볼륨 렌더링을 통해 임의의 시점에서 고품질 이미지를 생성한다. 이 두 차세대 실감미디어 기술을 통해 제한된 수의 입력 이미지만으로도 사실적인 자유 시점 뷰 합성이 가능하게 되었다. 그 중에서도 NeRF는 Implicit Neural Representation (INR)을 기반으로 하는데, INR은 연속적인 신호나 기하학적 형태를 신경망의 가중치로 암시적으로 표현하는 방식이며, 이러한 접근 방법은 SIREN[6]에서 주기적 함수를 도입하며 본격적으로 주목받기 시작했다. 이러한 INR은 전통적인 실감 미디어 처리 기법 중 하나인 라이트필드에도 적용되어 뉴럴 라이트필드[7]라는 새로운 접근 방식을 창출해 냈으며, 이는 기존 라이트필드의 데이터 취득의 어려움을 상당 부분 극복하는데 기여하였다.

라이트필드[8]는 자유 공간을 통과하는 무수히 많은 수의 광선 (light ray)을 정의하는 개념으로, 이 광선들을 모두 확보하였다는 가정 하에 해당 광선들의 조합을 바탕으로 원하는 시점의 다양한 뷰를 합성해 내는 기술이다. 3D 모델링[9,10], Image-based Rendering (IBR)[11,12,13]에서 요구되는 3D 지오메트리 추정이 불필요하다는 점에서 복잡도가 낮다는 큰 장점을 가진다. 하지만, 라이트필드는 자유 공간을 통과하는 모든 광선을 취득하고 보유해야 한다는 제약이 있으며, 이는 실제 구현에서 다시점 이미지로 대체되지만, 상당히 조밀하고 정교하게 배치된 다중 카메라 배치가 요구된다는 점에서 어려움이 크다. 뉴럴 라이트필드는 INR과 유사한 방식으로 라이트필드의 4차원 좌표 (u, v, s, t)를 입력으로 대응되는 (R, G, B) 값을 추정하는 MLP 네트워크를 학습한다. 이 과정에서, 상대적으로 간격이 넓고, 자유롭게 배치된 카메라에서 촬영된 이미지로도 충분히 학습이 가능하여, 기존 라이트필드의 데이터 취득의 어려움을 상당 부분 개선한다. 더불어, 뉴럴 라이트필드는 NeRF와 달리 볼륨 렌더링 과정 없이 광선에 따라 예측된 RGB를 그대로 사용하기 때문에 직관적이고, 효율성이 높은 장점 또한 가진다.

뉴럴 라이트필드는 기존 라이트필드에서 요구되는 다수의 이미지를 단 하나의 MLP로 줄였다는 점에서 상당히 큰 효율 개선을 달성한다. 하지만, XR 등과 같은 모바일 디바이스에서 활용되고, 고품질의 뷰 렌더링을 위해서는 더 큰 크기의 MLP 네트워크가 요구된다는 점에서 추가적인 최적화가 요구된다. 본 연구는 이러한 motivation을 바탕으로 NeuLF[14] 기반으로 학습된 뉴럴 라이트필드의 MLP 모델의 효율성 개선을 위해 은닉층 별 비트 정밀도 변화에 따른 성능 변화를 분석하고, 적절한 비트 정밀도 변환 방법을 제시한다.

a) 서울과학기술대학교 스마트ICT융합공학과(Dept. of Smart ICT Convergence Engineering, Seoul National University of Science and Technology)
‡ Corresponding Author : 정현민(Hyunmin Jung)
E-mail: hmjung@seoultech.ac.kr
Tel: +82-2-970-6457
ORCID: https://orcid.org/0000-0001-8216-5842
※이 논문의 결과 중 일부는 한국방송·미디어공학회 2025년 하계학술대회에서 발표한 바 있음
※This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00229330, 3D Digital Media Streaming Service Technology)
· Manuscript September 4, 2025; Revised October 15, 2025; Accepted October 15, 2025.

Ⅱ. 관련 연구

1. 라이트필드와 뉴럴 라이트필드

라이트필드는 자유 공간을 통과하는 모든 광선을 정의하고, 이의 조합을 통해 다양한 novel 뷰를 합성하는 기술이다[8]. 일반적으로 라이트필드는 하나의 광선을 표현하는데 (u, v, s, t)의 네 개 변수를 사용한다. 이를 위해 자유 공간 상에 평행하는 두 평면을 가정하며, 임의의 광선은 두 평면 위의 두 점, (u, v)와 (s, t)를 통해 해당 광선을 두 점을 이은 것으로 정의한다. 사용자의 요구 시점에 따라 필요한 광선들을 재조합함으로써 원하는 시점의 뷰를 합성해 낸다. 라이트필드의 광선은 실제 구현의 관점에서 이미지의 픽셀로 대체된다.

뉴럴 라이트필드[7]는 이러한 전통적인 라이트필드 표현에 INR[6] 기법을 적용하여 연속적이고 고해상도의 신호를 좌표 공간 위의 함수로 모델링한다. 이렇게 학습된 MLP는 고해상도 샘플링이 가능하면서도 메모리 오버헤드는 좌표-값 대응을 위한 가중치 수에 의존하기 때문에 고정적이다. 뉴럴 라이트필드는 4차원 라이트필드 좌표 (u, v, s, t)를 입력으로 받아 MLP를 통해 해당 광선의 색상 값을 예측한다.

뉴럴 라이트필드의 분야에서는 효율성과 성능 향상을 위한 다양한 연구가 진행되고 있다. NeuLF[14]는 2면 파라미터화를 통해 라이트필드를 효율적으로 표현하고, 볼륨 샘플링 없이 광선 당 한 번의 패스로 빠른 렌더링을 가능하게 했다. LightSpeed[15]는 2D Voxel Grid를 활용하여 모바일 환경에서의 실시간 뷰 합성을 실용화하는 데 기여하였다. 최근에는 N차원 Voxel Grid를 활용하여 뉴럴 라이트필드의 효율성과 정확도를 최적화한 연구와 함께[16], 기본 MLP 공유와 함께 보조 계층의 추가를 통해 확장 가능한 라이트필드 모델을 제시한 연구[17] 등 여러 연구들이 이어지고 있다.

2. 모델 경량화

딥러닝 네트워크 모델의 경량화는 제한된 자원에서의 딥러닝 활용을 위해 필수적인 기술이다. 모델 경량화는 중요도가 상대적으로 낮은 가중치를 제거해 네트워크 크기와 연산량을 줄이는 pruning과, 대형 teacher 모델 네트워크를 작은 student 모델 네트워크에 지식을 증류하는 지식 증류 (knowledge distillation) 등의 기법들이 사용된다. 뉴럴 라이트필드의 유사 기술인 NeRF 역시 다양한 경량화 방법이 시도되었다. VQRF[18]는 볼륨 radiance field 압축을 위해 voxel에 pruning 기법을 사용하며 100배 압축률을 달성한 바 있으며, R2L[19]은 사전 훈련된 대형 NeRF 모델을 지식 증류 기법을 통해 더 효율적인 모델로 증류한 바 있다.

비트 정밀도 조절은 부동소수점의 네트워크 가중치를 저비트로 근사하여 모델의 메모리 및 연산 비용을 줄이는 대표적인 경량화 기법이다. HNeRV[20]는 CNN 기반의 모델에 8bit로의 정밀도 조절을 사용하여 모델 크기를 75% 압축하면서도 PSNR 성능 저하를 1dB 이내로 억제한 바 있다. 최근 연구에서는 MLP 기반 비전 모델에서 저비트 변환 과정의 문제점을 체계적으로 분석하고, 활성화 범위 제한과 민감성 레이어에 대한 기법을 도입해 4bit, 8bit 저비트 변환에서도 성능 저하를 최소화하는 방법들이 제안되었다[21].

이들은 모델의 정확도 손실을 최소화면서도 실질적인 성능 향상을 제공하며, 특히, 실시간 추론이 요구되는 응용 분야에서 배포 가능한 수준의 모델을 만드는 데 핵심적인 역할을 한다. 각 기법은 정확도와 효율성 간의 균형점을 찾는 것이 중요하며, 목표 환경의 제약 조건과 요구 성능에 따라 적절한 기법을 선택하거나 조합하여 최적화를 수행해야 한다.

Ⅲ. 제안 방법

1. 4D 라이트필드 표현과 뉴럴 라이트필드

4D 라이트필드는 자유 공간을 통과하는 광선을 4개의 변수를 통해 표현하는 방법을 나타내며, 그림 1 (a)는 이를 그림과 함께 소개한다. 자유 공간 상에 평행하는 두 평면을 가정하며, 해당 평면을 통과하는 광선을 두 평면을 통과하는 두 개의 점, (u, v)와 (s, t)로 정의한다. 이를 통해 두 평면을 통과하는 모든 광선을 정의한다. 그림 1 (b)는 이처럼 정의된 광선에 대하여 사용자의 임의의 점에 따른 뷰를 생

성하는 과정을 보여준다. 해당 뷰를 표현하기 위해 필요한 광선들을 정의하고, 해당 광선에 대응되는 R, G, B를 이미지 형태로 재구성한다. 임의의 시점 뷰 합성에 있어 3D 지오메트리 추정과 같은 복잡한 과정 없이, 광선의 재조합이라는 단순한 과정만 요구된다는 점에서 장점을 가진다.

뉴럴 라이트필드는 이러한 광선의 (u, v, s, t) 좌표계와, R, G, B의 색 정보 사이의 관계를 MLP를 통해 학습한다. 그림 1 (c)는 뉴럴 라이트필드의 MLP 모델과 입, 출력을 보여준다. 예측한 R, G, B 값인 cpred와 실제 ground truth인 c 사이의 차이를 바탕으로 MLP를 학습하며, 식 (1)과 같은 광도 손실(Photometric Loss) 함수를 사용한다.

$$L_p = \sum_{r \in R} \| c_{pred} - c \|_2 \tag{1}$$

본 연구에서 사용한 뉴럴 라이트필드의 MLP 모델은 은닉층의 개수가 8, 각 은닉층의 요소 수가 512인 구조를 사용한다. 두 개의 입, 출력 layer를 포함하며, 또한 skip connection은 6번째 은닉층의 입력과 연결되도록 설계되었다.

2. 뉴럴 라이트필드 MLP 네트워크의 비트 정밀도 조정

본 연구에서는 그림 1 (c)의 뉴럴 라이트필드 MLP 네트워크를 대상으로 비트 정밀도를 기존 32bit에서 16bit 또는 8bit로 근사함에 따른 화질 대비 네트워크 크기 부담 사이의 효율을 분석한다. 본 연구에서는 사전 학습된 MLP 네트워크를 대상으로 비트 정밀도 변환을 적용하며, 비트 정밀도 변환을 적용하는 은닉층에 따라 성능에 미치는 민감도를 함께 분석한다. 본 연구에서는 각 뉴런에 대하여 은닉층별 min-max 범위를 이용한 균일 선형 방식으로 비트 정밀도를 조정하며, 이는 식 (2)와 (3)과 같다.

$$u_i' = \mathrm{Round}\left(\frac{u_i - u_{\min}}{s}\right) * s + u_{\min} \tag{2}$$

$$s = \frac{u_{\max} - u_{\min}}{2^b - 1} \tag{3}$$

식 (2)와 (3)에서 ui는 은닉층의 뉴런의 가중치를 의미하며, umax와 umin는 해당 은닉층의 최대, 최소 가중치를 의미한다. Round는 반올림 함수를, b는 비트 정밀도로 변화하는 bit 크기를 의미한다. s는 가중치 값의 범위에 대한 인수이다. 가중치의 복원을 위해서는 bit 크기가 조정된 뉴런과 함께 s와 umin이 저장되어야 한다.

표 1은 은닉층의 가중치의 비트 정밀도를 기존 32bit에서 8bit로 조정함에 따른 MLP 네트워크의 모델 크기 변화를 보여준다. Index는 비트 정밀도 변환을 적용한 은닉층의 번호를 의미한다. 비트 정밀도 조정이 적용되지 않은 기존 모델의 네트워크 크기는 9.1MB이다. 8번째 은닉층을 대상으로 비트 정밀도를 8bit로 조정할 경우 네트워크 크기는 8.3MB로 감소한다. 7, 6, 5, 4, 3, 2, 1번째 은닉층으로 적용 범위를 넓혀 감에 따라 약 0.8MB의 크기가 고정적으로 감소한다. 다만 6번째 layer의 경우 skip connection으로 인해

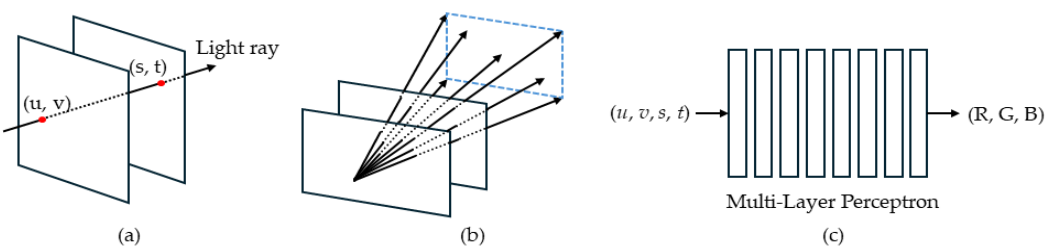

그림 1. 뉴럴 라이트필드 소개 (a) 광선을 정의하는 네 개 변수 (b) 임의의 시점에 대한 뷰 생성을 위한 광선 조합 (c) 뉴럴 라이트필드 학습을 위한 MLP 네트워크

Fig. 1. Introduction of Neural Light Field (a) The four variables defining a ray (b) Combination of rays for generating a arbitrary view (c) MLP Network for Neural Light Field

표 1. 뉴럴 라이트필드 MLP 모델의 은닉층에 대하여 비트 정밀도를 8bit로 조정함에 따른 모델 크기 변화
Table 1. Changes in model size from adjusting the 8bit precision for the hidden laeyr of the Neural Light Field MLP model

Index	MLP Network Size
-	9.1MB
8	8.3MB
8, 7	7.6MB
8, 7, 6	6.1MB
8, 7, 6, 5	5.3MB
8, 7, 6, 5, 4	4.6MB
8, 7, 6, 5, 4, 3	3.9MB
8, 7, 6, 5, 4, 3, 2	3.1MB
8, 7, 6, 5, 4, 3, 2, 1	2.3MB

약 1.5MB의 상대적으로 큰 감소가 발생한다. 모든 은닉층에 비트 정밀도 변환을 적용 시 모델 크기가 2.3MB로 압축되며 약 25%로의 네트워크 크기를 줄일 수 있음을 볼 수 있다.

Ⅳ. 실험 결과

1. 실험 환경

본 논문은 비트 정밀도 조정에 따른 네트워크 모델 크기와 화질 저하 사이의 효율을 분석하기 위해 Stanford Light Field Dataset[22]로 학습된 뉴럴 라이트필드 모델을 사용한다. Stanford Light Field Dataset는 정교하게 배치된 카메라를 사용하여 촬영된 4D 라이트필드 데이터세트이며, 총 17×17의 viewpoint로 구성된다.

뉴럴 라이트필드 MLP 모델의 학습은 5e-4의 초기 learning rate을 기반으로, 매 에폭마다 0.995만큼 감소하도록 스케쥴링을 적용하였으며, 총 1K 에폭의 학습을 수행한 결과를 사용한다. 학습이 완료된 MLP 네트워크 모델에 대하여 은닉층 단위로 16bit 또는 8bit의 비트 정밀도 조정을 적용한다.

2. 일괄적인 Bit Precision 조정에 따른 분석

표 2는 MLP 전체 레이어에 대해 16bit, 8bit로의 비트 정밀도 조정을 일괄적으로 적용한 모델의 화질 저하를 비교한다. 비트 정밀도 조정을 모든 은닉층에 16bit로 적용한 결과, PSNR은 평균 39.63dB에서 39.62dB로 단 0.01dB의 미미한 성능 저하를 보인다. 각 샘플별 비교에서도 거의 유사한 PSNR 결과를 보인다. 반면, 8bit 정밀도로 일괄 조정 시, PSNR은 평균 28.27dB로 크게 하락한 결과를 보인다.

표 2. 전체 layer에 대한 비트 정밀도 변환의 일괄 적용에 따른 화질 저하 비교
Table 2. Comparision of image quality degradation from adjusting bit precision to the entrie layer

Sample	Original (9.1MB)	16bit precision (4.6MB)	8bit precision (2.3MB)
beans	43.75	43.75	40.98
bracelet	39.93	39.93	27.91
bulldozer	41.34	41.34	28.78
bunny	44.74	44.73	36.96
chess	42.74	42.73	33.81
flowers	38.92	38.92	26.27
gem	43.09	43.09	33.51
knights	37.30	37.32	25.55
tarot	27.00	27.01	20.15
tarot_small	38.26	38.25	21.50
treasure	35.46	35.37	12.28
truck	43.05	43.05	31.50
average	39.63	39.62	28.27

그림 2는 표 2의 비트 정밀도 일괄 조정에 따른 화질 변화를 질적으로 비교한다. 그림 2는 treasure와 bulldozer 두 개 샘플에 대한 결과를 비교하며, 가장 왼쪽에서부터 ground truth, 32bit 원본 모델, 16bit 정밀도 일괄 적용 결과, 8bit 정밀도 일괄 적용 결과를 보여준다. 16bit 정밀도 변환의 결과를 보면, 표 2의 결과에서 PSNR 변화가 거의 없었던 것과 마찬가지로, Original의 결과와 거의 동일함을 보여준다. 반면, 8bit 정밀도 변환 적용 결과에서는 큰 화질 저하를 보인다. 특히 treasure 샘플의 경우 대상 샘플을 거의 표현하지 못하는 결과를 보여준다. Bulldozer 샘플의 경우 대상이 잘 표현된 결과를 보여주지만 확대된 그림에서 왜곡된 부분을 확인할 수 있다.

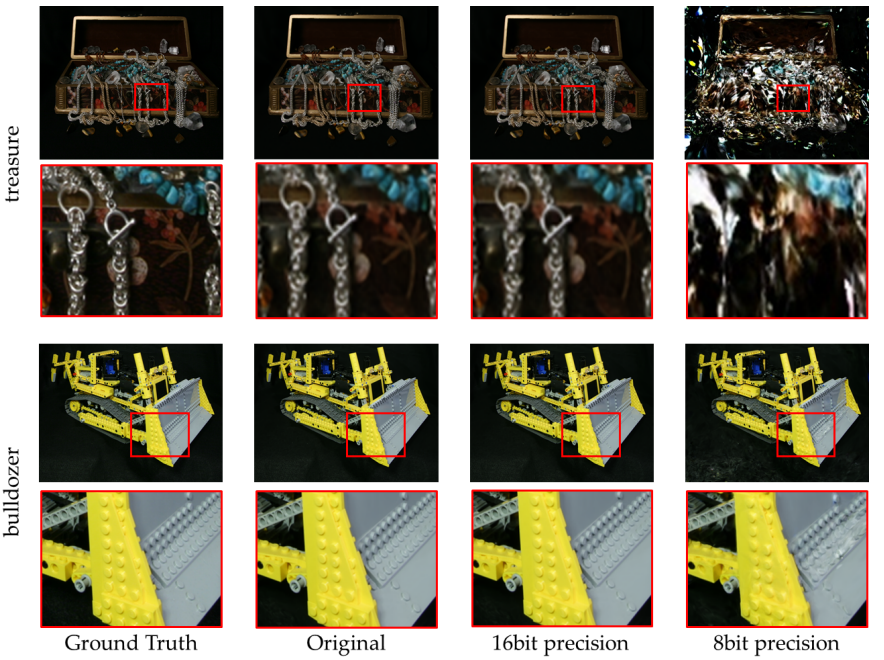

그림 2. 전체 은닉층에 대한 비트 정밀도의 일괄 조정에 따른 화질 변화의 질적 비교
Fig. 2. Qualitative comparison of image quality changes resulting from applying bit precision to the entire hidden layer

표 3. 개별 은닉층에 대한 8bit 정밀도 변환 적용에 따른 평균 PSNR 비교
Table 3. Comparision of average PSNR with 8bit precision applied to individual hidden layers

Index	1	2	3	4	5	6	7	8	Original
PSNR(dB)	30.29	30.88	35.81	37.58	37.97	38.23	38.01	37.67	39.63

3. 개별 은닉층 단위의 세부 조정에 따른 결과 비교

표 2와 그림 2에서 확인할 수 있듯이, 16bit 정밀도 일괄 조정은 거의 화질 저하가 없으면서 모델 크기를 절반으로 줄이는 결과를 보인다. 반면, 8bit 정밀도 조정은 모델 크기 측면의 더 큰 감소 효과를 보이지만, 눈에 띄는 화질 저하로 적용이 어렵다. 이에, 우리는 일부 은닉층을 대상으로 8bit 정밀도 조정을 적용하여 세부적으로 분석한다. 먼저, 표 3은 8개의 은닉층들을 각각 8bit 정밀도로 조정함에 따른 평균 PSNR 변화를 보여준다. 표 3의 결과를 보면, 1번째 은닉층에 대한 8bit 정밀도 조정 시 30.29dB로 큰 폭의 화질 저하를 보이는 반면, 8번째 은닉층에 대한 적용은 37.67dB로 상대적으로 적은 저하를 보인다. 전반적으로 output layer에 가까운 4, 5, 6, 7, 8번째 은닉층에 대한 비트 정밀도 조정에서 민감도가 낮은 결과를 보인다. 특히, input layer에 가까운 1, 2번째 은닉층에 대한 조정은 30dB 대의 굉장히 낮은 결과를 보인다.

이는 앞쪽 레이어에 대한 비트 정밀도 감소가 해당 구간의 계산 오류를 연쇄적으로 누적시켜 이후 단계로 전이되기 때문으로 해석할 수 있다. 반면, 출력 계층에 가까운 후반부 은닉층의 경우, 네트워크 내부에서 비교적 안정된 표현이 형성된 상태에서 양자화가 적용되므로, 입력 신호 변형에 대한 민감도가 낮아진다. 이러한 관점은 MLP 신경망에서 계층별 표현 안정성과 비트 정밀도 감소로 인한 오류 전파 특성을 고려한 비트 정밀도의 최적화 전략을 설계하는데 중요한 시사점이라고 할 수 있다. 따라서 이러한 후반부 은닉층에 대해 비트 정밀도 변환을 조정하는 전략은 비교적 일반적으로

적용 가능한 최적화 전략이라고 볼 수 있다.

4. Output Layer에 가까운 은닉층의 비트 정밀도 조정

위 실험 결과를 바탕으로 output layer에 가까운 8번째 은닉층부터 점진적으로 비트 정밀도 변환을 적용하는 전략이 성능 저하를 최소화할 수 있다는 계획을 세웠으며, 그림 3의 그래프는 이를 모든 샘플에 적용한 결과를 보여준다. 이는 8번째 은닉층부터 비트 정밀도 변환 범위를 점차적으로 늘려감에 따라 PSNR이 점차적으로 감소하는 결과를 보여준다. 대부분의 샘플에서 3번째 은닉층까지 PSNR 저하가 크지 않았으며, 2번째, 1번째 은닉층을 포함 시 PSNR이 급격히 저하하는 결과를 보인다. 표 1에서 살펴본 바와 같이 8번째에서 3번째 은닉층까지의 8bit 정밀도 변환 적용 시 모델 크기는 3.9MB로 기존 9.1MB 대비 큰 네트워크 절감 효과를 보인다.

Treasure와 flowers 샘플의 경우 비교적 빠른 PSNR 저하를 보이는데, treasure 샘플의 경우 8번째 은닉층에서의 적용에서 가장 큰 PSNR 저하를 보였으며, 2번째 은닉층 적용 시 20dB 이하의 굉장히 낮은 PSNR 결과를 보였다. Flowers 샘플의 경우 마찬가지로 8번째 은닉층에서의 적용 시 큰 PSNR 저하를 보였으며, 이후에는 유사한 수준의 저하를 보였다. 각 샘플 모델에 대한 비트 정밀도 변환 결과를 비교해보았을 때, 원본 모델의 PSNR이 낮은 샘플로 학습된 모델일수록 비트 정밀도 변환에 더 민감하게 반응하는 경향을 보였다. 이는 네트워크가 학습하기 어려운 데이터를 기반으로 한 모델이 비트 정밀도 감소에 더 취약한 특징을 보인다고 해석할 수 있다.

그림 4는 이처럼 은닉층에 순차적으로 비트 정밀도 변환을 적용함에 따른 렌더링 결과를 질적으로 비교한다. 해당 결과는 treasure 샘플에 대한 결과이며, 앞선 그림 2의 일괄 적용 시 큰 폭의 화질 저하를 보인 샘플에 해당된다. 그림 4의 결과를 보면, 8번째 은닉층에서 3번째 은닉층까지의 비트 정밀도 변환 적용 결과까지는 비교적 원본과 유사한 형태를 보임을 확인할 수 있으며, 다만, 색감의 차이가 확인된다. 2번째, 1번째 은닉층까지 적용 시 급격한 화질 저하가 확인된다.

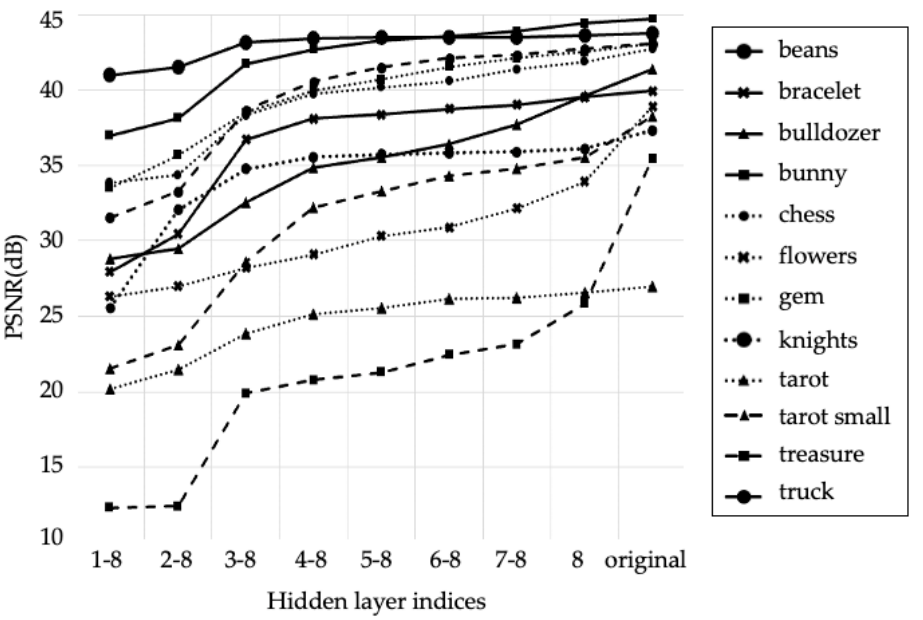

그림 3. 은닉층에 8bit 정밀도 변환을 점차적으로 조정함에 따른 PSNR 변화
Fig. 3. PSNR changes from gradually adjusting the 8bit precision in the hidden layer

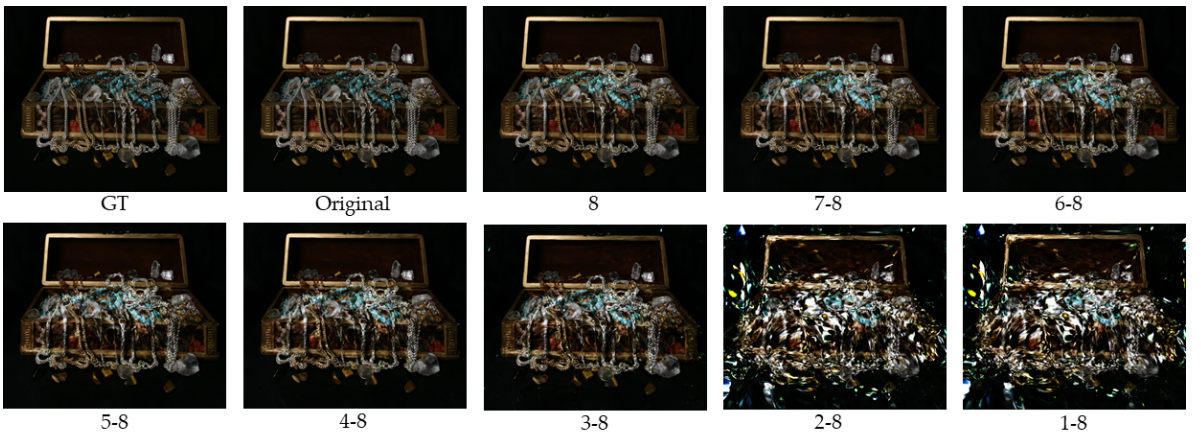

그림 4. Output layer와 가까운 은닉층에 대한 비트 정밀도 변환 적용 범위를 점진적으로 늘려감에 따른 화질 변화의 비교
Fig. 4. Comparison of image quality changes as the bit precision adjustment range for hidden layers near the output layer is gradually increased

Ⅴ. 결론 및 향후 계획

본 연구는 모바일 환경에서 뉴럴 라이트필드의 실용적 활용을 위해, 비트 정밀도 조절에 따른 모델 절감 효과와 화질 저하를 분석한다. 실험 결과, 비트 정밀도를 원본 32bit에서 16bit로 전체 은닉층에 대해 일괄적으로 조정할 경우 모델 크기를 절반으로 줄일 수 있으면서 PSNR 손실이 0.01dB로 거의 저하가 발생하지 않음을 확인할 수 있었다. 반면, 추가적인 모델 절감 효과를 위해 8bit로 일괄 조정할 경우 절감 효과가 두 배로 증가하는 반면, PSNR이 11.36dB로 낮아지는 심각한 화질 저하가 발생하였다. 이를 개선하기 위해 우리는 은닉층의 위치에 따른 민감도를 분석하고, output layer와 가까운 은닉층에 대하여 선택적으로 8bit 정밀도로 조정하는 전략을 제안하였으며, 이 방법을 통해 16bit 모델 대비 추가적인 절감 효과를 누리면서도 8bit 모델 대비 심각한 화질 저하를 피할 수 있었다. 향후 연구에서는 본 논문의 최적화 모델을 실제로 XR 디바이스에 적용해 봄으로써 성능을 검증하고, 추가적인 최적화 방안의 연구를 계획한다.

참 고 문 헌 (References)

[1] Meta Quest, https://www.meta.com/quest, (accessed Sep. 4, 2025).

[2] Apple Vision Pro, https://support.apple.com/us-en/117810, (accessed Sep. 4, 2025).

[3] A. Tewari et al., “State of The Art on Neural Rendering,” In Computer graphics forum, Norrköping, Sweden, Vol.39, No.2, pp.701-727, May, 2020.
doi: https://doi.org/10.1111/cgf.14022

[4] B. Kerbl et al., “3D Gaussian splatting for real-time radiance field rendering,” ACM Trans. Graph., Vol.424, No,139, pp.1-14, 2023.
doi: https://doi.org/10.1145/3592433

[5] B. Mildenhall et al., “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” Communications of the ACM, 65, pp. 99-106.
doi: https://doi.org/10.1145/3503250

[6] V. Sitzmann et al., “Implicit Neural Representations with Periodic Activation Functions,” Advances in neural information processing systems, Vol.33, pp.7462-7473, 2020.

[7] B. Chen, L. Ruan, and M. L. Lam, “LFGAN: 4D light field synthesis from a single RGB image,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol.16, No.1, pp.1-20, 2020.
doi: https://doi.org/10.1145/3366371

[8] S. J. Gortler et al., “The lumigraph,” In Seminal Graphics Papers: Pushing the Boundaries, Vol.2, pp.453-464, 2023.
doi: https://doi.org/10.1145/237170.237200

[9] F. Remondino, and S. El-Hakim, “Image based 3D modelling: a review,” The photogrammetric record, Vol.21, No.115, pp.269-291, 2006.
doi: https://doi.org/10.1111/j.1477-9730.2006.00383.x

[10] M. Livesu et al., “From 3D models to 3D prints: an overview of the processing pipeline,” In Computer Graphics Forum, Vol.36, No.2, pp.537-564, May, 2017.
doi: https://doi.org/10.1111/cgf.13147

[11] L. McMillan, and G. Bishop, “Plenoptic modeling: An image-based rendering system,” In Seminal Graphics Papers: Pushing the Boundaries, Vol.2, pp.433-440, 2023.
doi: https://doi.org/10.1145/3596711.3596758
[12] H. Shum, and S. B. Kang, “Review of image-based rendering techniques,” Visual Communications and Image Processing 2000, Perth, Australia, Vol.4067, pp.2-13, 2000.
doi: https://doi.org/10.1117/12.386541
[13] C. Jung, and L. Jiao, “Reliable depth-image-based rendering using parameter approximation in mobile devices,“ IEICE Electronics Express, Vol.7, No.10, pp.666-671, 2010.
doi: https://doi.org/10.1587/elex.7.666
[14] Z. Li et al., “NeuLF: Efficient Novel View Synthesis with Neural 4D Light Field,” EGSR (ST), 2022.
doi: https://doi.org/10.2312/sr.20221156
[15] A. Gupta et al., “LightSpeed: light and fast neural light fields on mobile devices,” Advances in Neural Information Processing Systems, Vol.36, No.1352, pp.31021-31037, 2023.
[16] I. G. Jeong, and H. Jung, “Neural light fields with N-dimensional voxel grids: a performance evaluation across voxel grid dimension,” IEICE Electronics Express, Vol.22, No.9, 20250141-20250141, 2025.
doi: https://doi.org/10.1587/elex.22.20250141
[17] I. G. Jeong, and H. Jung, “Scalable Neural Light Field with Layer Add-ons of Multi-Layer Perceptron,” IEEE MultiMedia, 2025.
doi: https://doi.org/10.1109/MMUL.2025.3581588
[18] L. Li et al., “Compressing volumetric radiance fields to 1 mb,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, pp.4222-4231, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.00411
[19] H. Wang et al., “R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis,” In European Conference on Computer Vision, Tel Aviv, Israel, pp.612-629, 2022.
doi: https://doi.org/10.1007/978-3-031-19821-2_35
[20] H. Chen et al., “HNeRV: A Hybrid Neural Representation for Videos,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, pp.10270-10279, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.00990
[21] L. Zhao, Z. Dong and K. Keutzer, “Analysis of quantization on mlp-based vision models,” arXiv preprint arXiv:2209.06383, 2022. (Accessed: Sep. 4, 2025).
[22] B. Wilburn et al., “High performance imaging using large camera arrays,“ In ACM siggraph 2005 papers, pp.765-776, 2005.
doi: https://doi.org/10.1145/1073204.1073259

저 자 소 개

양 서 준

- 2025년 2월 : 서울과학기술대학교 전자IT미디어공학과 학사
- 2025년 3월 ~ 현재 : 서울과학기술대학교 스마트ICT융합공학과 석사 과정
- ORCID : https://orcid.org/0009-0007-4765-7156
- 주관심분야 : 영상처리, 컴퓨터 비전

정 현 민

- 2014년 2월 : 경희대학교 전자전파공학과 학사
- 2016년 2월 : 서울대학교 전기정보공학부 석사
- 2020년 8월 : 서울대학교 전기정보공학부 박사
- 2023년 3월 ~ 현재 : 서울과학기술대학교 스마트ICT융합공학과 조교수
- ORCID : https://orcid.org/0000-0001-8216-5842
- 주관심분야 : 실감미디어(VR, AR, XR, 메타버스), 영상처리, 컴퓨터 비전